Project_Report

```r
## Step 1 : load data

library(readr)
collegedata <- read_csv("MERGED2000_01_PP.csv") # (I renamed it collegedata)
dim(collegedata) # output: number of rows, number of cols
```

```
## [1] 6654 1986
```

```r
goodstuff_UG = collegedata$UG != "NULL"  # identifying all non NULL UG rows
collegedata_clean = collegedata[goodstuff_UG, ] # Keeping only non NULL  UG rows
goodstuff_DEBT = collegedata_clean$DEBT_MDN != "NULL" # Identifying all non UG DEBT_MDN rows
collegedata_clean_2 = collegedata_clean[goodstuff_DEBT, ] # Keeping only non NULL DEBT_MDN rows

goodstuff_DEBT_Privacy_rm = collegedata_clean_2$DEBT_MDN != "PrivacySuppressed" # identifying all non P
collegedata_use = collegedata_clean_2[goodstuff_DEBT_Privacy_rm, ] # keeping all of the non PrivacySupp
collegedata_use$UG = as.numeric(collegedata_use$UG) # changing values to numeric
collegedata_use$DEBT_MDN = as.numeric(collegedata_use$DEBT_MDN) # changing values to numeric
```

Description of Data:

I elected to use the U.S. Dept. of Education's College Scorecard dataset, because I thought it would be interesting to investigate the financial standing of colleges across the country. I think that trends in educational management (yearly tuition increases, bloating administrative costs and luxury facilities, increased alumni debt) may have weakened the longevity of colleges and universities as a whole, and the recent shock to the economy caused by the pandemic may well lead to the demise (or substantial restructuring) of a large swath of these institutions. The College Scorecard contains data about the student and alumni populations, and these are important markers of the longevity of institutions whose financial health is predominantly based on donations from these alumni.

The data was collected by the federal government via the census, Department of Education records, and IRS federal tax returns. Income and debt repayment data only include students who received federal financial aid (Pell grants, etc.). The dataset includes information going back multiple decades and is the largest dataset on higher education ever published.

The observational units for this dataset are individuals who attended college in the U.S. and received federal financial aid, as well as institutions of higher education which receive federal financial benefits from the U.S. government.

Students who did not receive federal financial aid are excluded from this dataset. This still leaves 46% of all students in the country within the dataset (Rothwell).

Since I attend a small college (Whitman is only about 1,500 students), and I receive federal financial aid, I was interested in learning if there was some relationship around this population level that perhaps did not exist elsewhere. This is why my analysis is performed at the small college level.

I analyzed the data provided in the Scorecard for all institutions in the U.S. excluding where data points were null or nonnumeric, i.e. "Privacy Suppressed" data for individuals who did not respond to surveys or had their information stricken from the Scorecard for any reason.

My variables of interest were the median original amount of the loan principal upon entering repayment of students (“DEBT_MDN”) and the total undergraduate population of students as of 2001 (“UG”). This is because the undergraduate population data is only available for the year 2001. My null hypothesis was that there was no correlative relationship between these two quantitative variables and my alternative was that any relationship existed. Expressed formally, in a linear regression model:

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

Summary Statistics:

```
## Summary statistics

summary(collegedata_use$UG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     152     734    2626    2740   46834 
```

```
summary(collegedata_use$DEBT_MDN)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     500    2825    4914    5647    6742   23300 
```

```
## Removing datapoint with 0 UG population

collegedata_use = collegedata_use[-which(collegedata_use$UG == 0), ]
```

Summary Stats and Visualization:

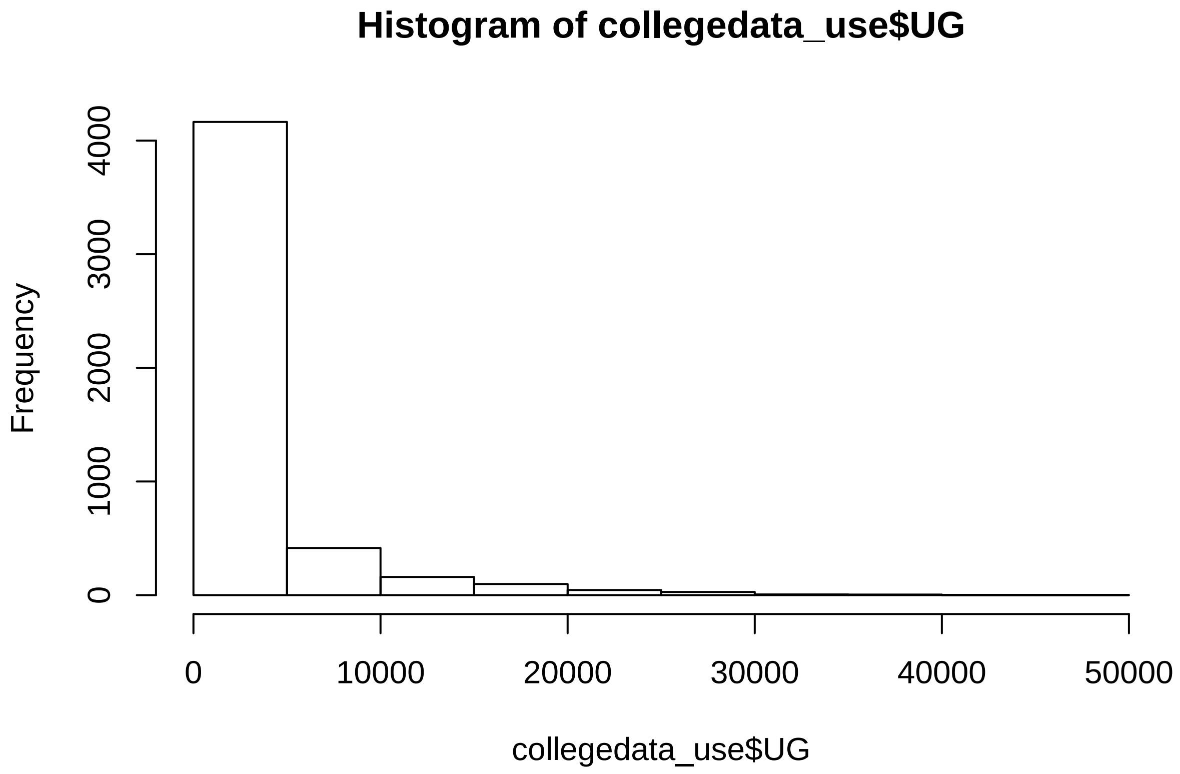
```
## Summary statistics and graphs

summary(collegedata_use$UG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       1     154     742    2635    2744   46834 
```

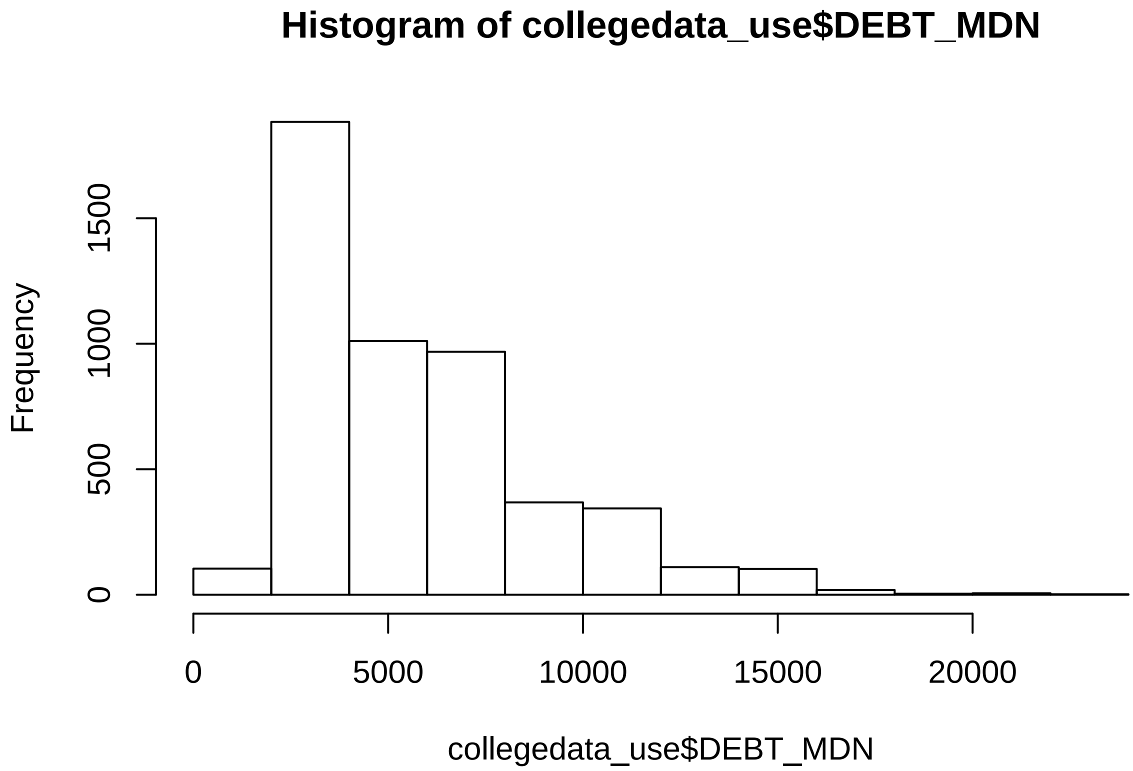
```
summary(collegedata_use$DEBT_MDN)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     500    2821    4913    5642    6735   23300 
```

```
hist(collegedata_use$UG)
```



```
hist(collegedata_use$DEBT_MDN)
```



```
#plot(collegedata_use$UG, collegedata_use$DEBT_MDN)
#plot(log(collegedata_use$UG), log(collegedata_use$DEBT_MDN))
```

Descriptive Stats for All Schools

From our histograms, we see that both of our variables of interest are unimodal, but they both differ greatly from the standard normal distribution. The undergraduate population has no left tail and the median original amount of the loan principal upon entering repayment has a very small left tail. A vast majority of institutions of higher education have undergraduate populations under 5,000.

Descriptive Stats for Small Schools Alone

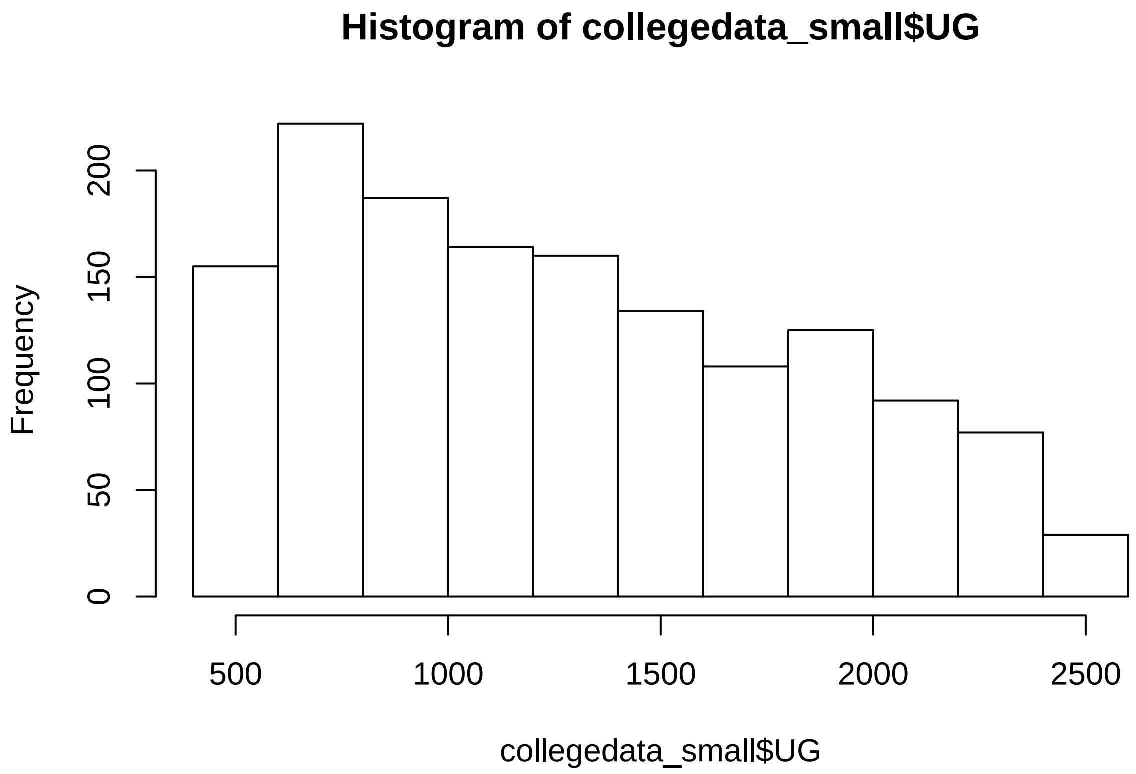
```
## Exploring only institutions with IG between 500 and 2500

collegedata_small = collegedata_use[collegedata_use$UG > 500 & collegedata_use$UG < 2500, ]

summary(collegedata_small$UG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     501     787    1199    1280    1719    2498
```

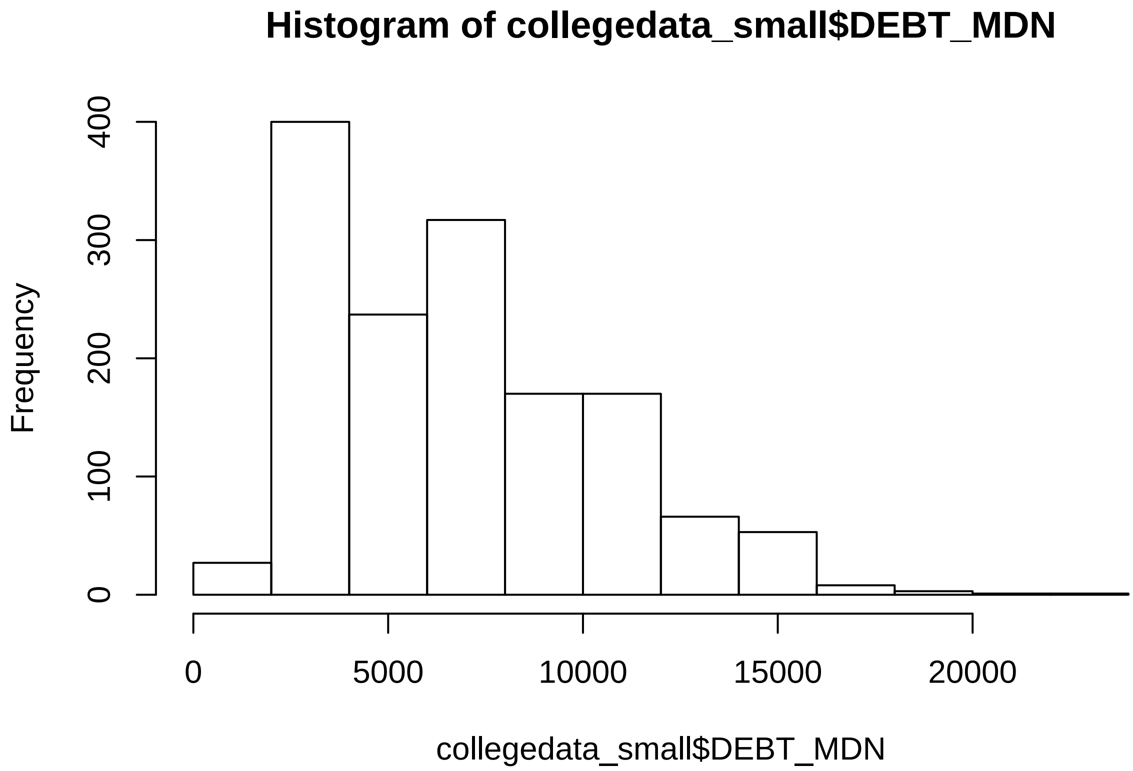
```
summary(collegedata_small$DEBT_MDN)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     867    3500    6125    6769    9250   23262
```

```
hist(collegedata_small$UG)
```



```
hist(collegedata_small$DEBT_MDN)
```



```
#plot(log(collegedata_small$UG), log(collegedata_small$DEBT_MDN))
```

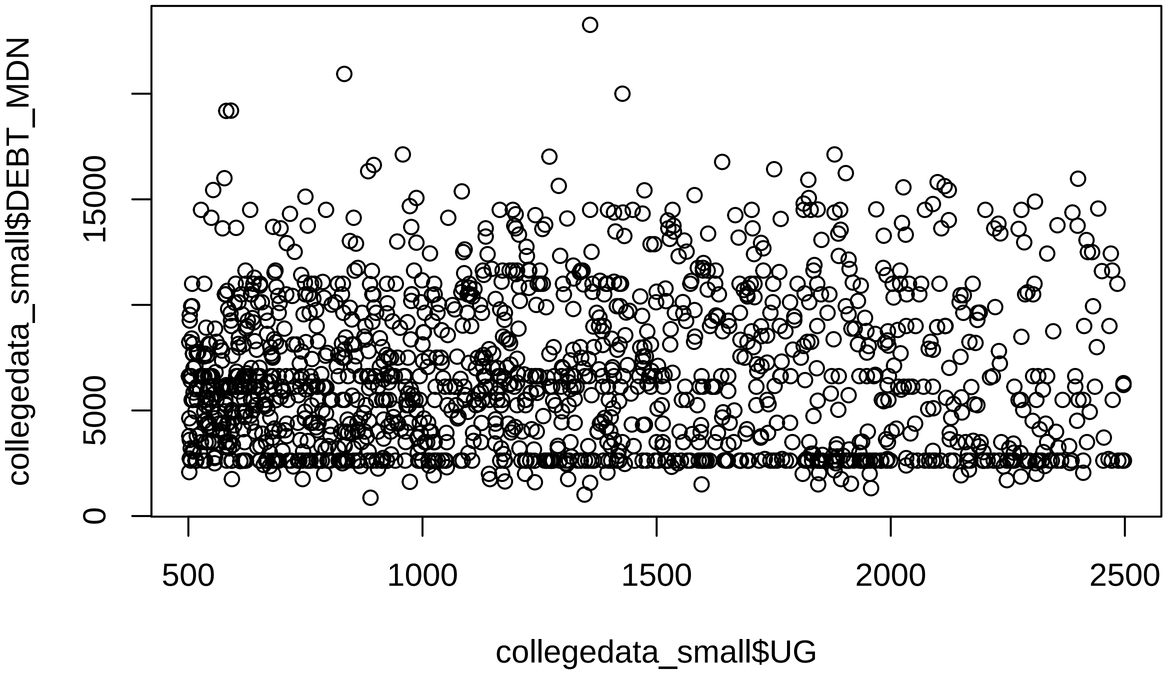
Our small-school-only histograms tell us the distribution of the variables for schools of undergraduate size 2500 and under. We see that the median debt amount is close to normal, but has a peak on the left and a very small tail, without a very smooth tail on the right. We can almost say this variable is bimodal. The undergraduate population peaks on the left as well, but its tail on the right is much smoother. In order for our data to be valid for analysis, we know that all values reported must be independent of each other. This is true in our case.

```
cor(collegedata_small$UG, collegedata_small$DEBT_MDN)
```

```
## [1] 0.03269607
```

We also see that the relationship between these two variables has a correlation (by cor()) of close to 0. This suggests there is not a linear relationship between these two variables. These facts do not bode well for our alternative hypothesis; we know that in order to obtain a positive result, we must have normally distributed data, and our variables must be related by linearity.

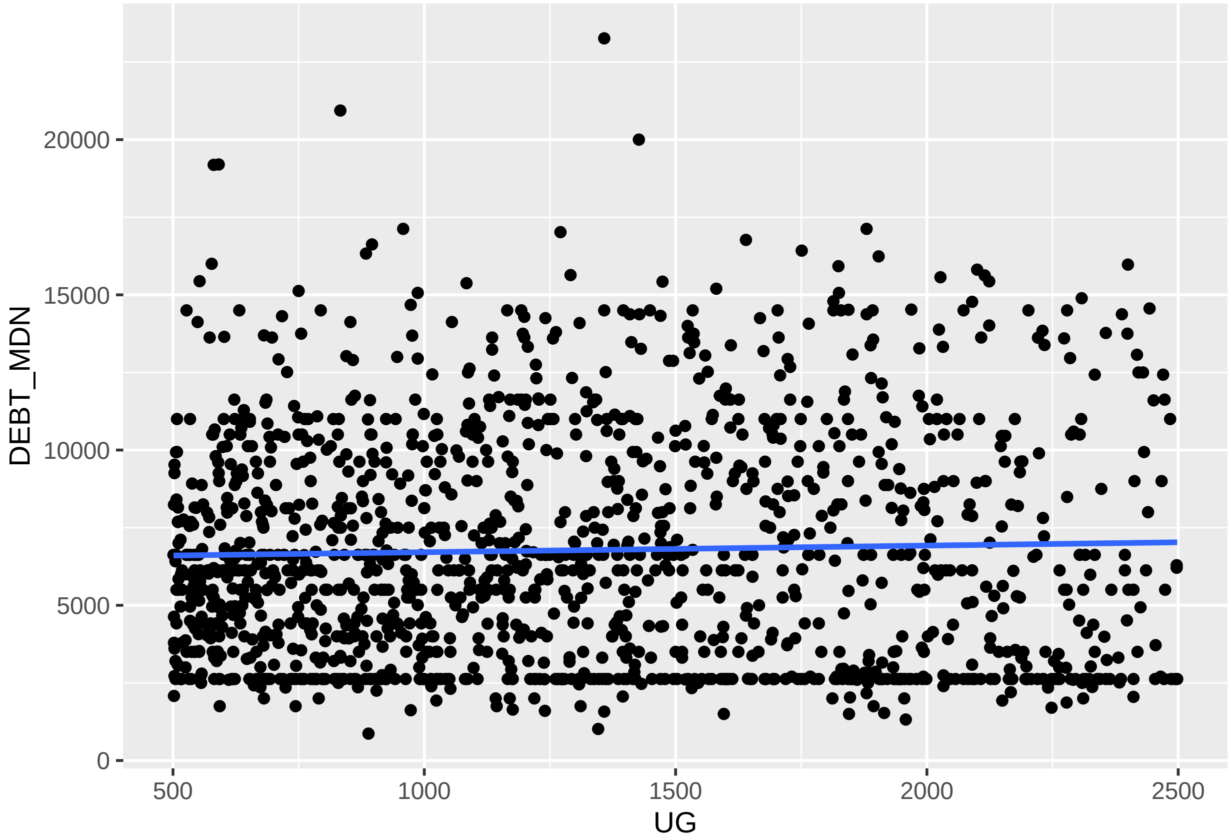
```
plot(collegedata_small$UG, collegedata_small$DEBT_MDN)
```



From our scatter plot we have more evidence that there seems to be no linear relationship between the median debt amount and the undergraduate population.

```
## Other helpful graphs
library(ggplot2)
#ggplot(data = collegedata_small, aes(x = UG, y = DEBT_MDN)) +
  #geom_point()

lmmodel <-lm(DEBT_MDN~sqrt(UG), data = collegedata_small)

ggplot(data = collegedata_small, aes(x = UG, y = DEBT_MDN)) +
  geom_point()+
  geom_smooth(method=lm, se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
lmmodel
```

```
##
## Call:
## lm(formula = DEBT_MDN ~ sqrt(UG), data = collegedata_small)
##
## Coefficients:
## (Intercept)     sqrt(UG)
##     6121.88        18.55
```

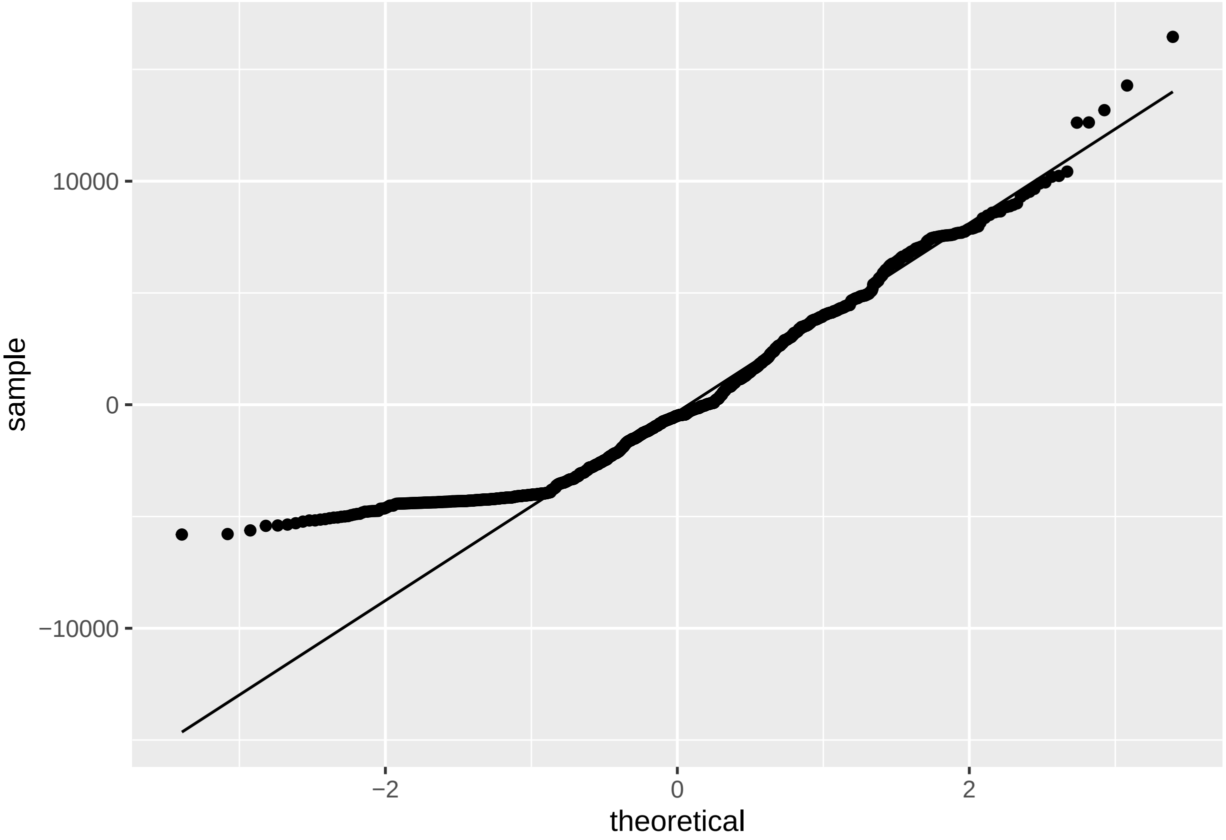
```
summary(lmmodel)
```

```
##
## Call:
## lm(formula = DEBT_MDN ~ sqrt(UG), data = collegedata_small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5807.8 -3168.9  -499.1  2523.4 16456.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6121.88     435.07  14.071   <2e-16 ***
## sqrt(UG)       18.55      12.16   1.525    0.127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3628 on 1451 degrees of freedom
## Multiple R-squared:  0.001601,   Adjusted R-squared:  0.0009125
## F-statistic: 2.326 on 1 and 1451 DF,  p-value: 0.1274
```

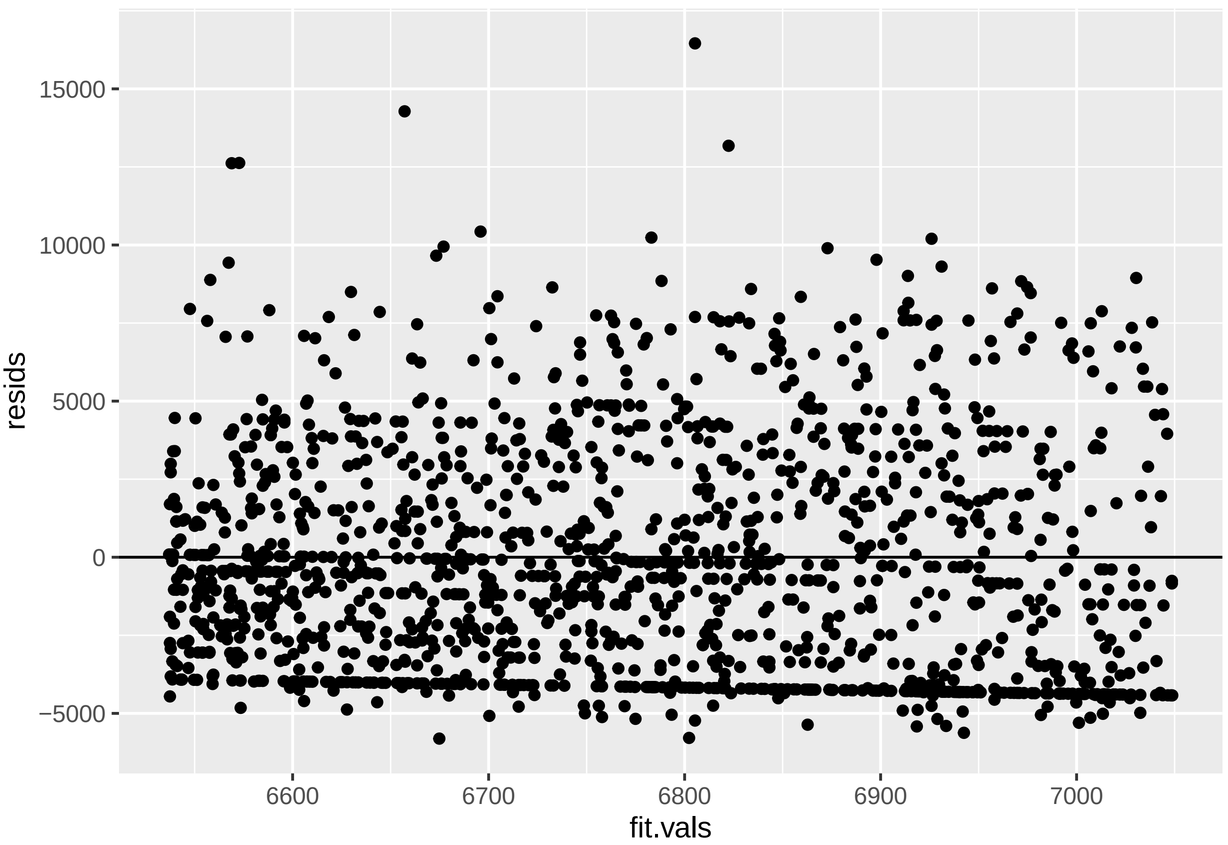
In our regression model, we can observe even without without using the cor() function that r is close to 0, which tells us that we have a very weak (if any) correlation between the undergraduate population of a college and the median debt amount of a student there. From the linear line we created, we found that the slope of our line with respect to the square root of the undergraduate population is 18.55, intersecting the y-axis at 6121.88. This tells us that our very weak correlation takes a positive slope.

```
resids <- lmmodel$residuals
fit.vals <- lmmodel$fitted.values
assess.fit <-data.frame(resids, fit.vals)
ggplot(assess.fit,aes(sample =resids))+
  stat_qq()+
  stat_qq_line()
```



Since our residuals QQ-plot follows the linear pattern fairly well except at the bottom, we can say it is normally distributed for values above -1 standard deviation from the theoretical mean.

```
ggplot(data = assess.fit)+
  geom_point(mapping =aes(x = fit.vals, y = resids))+
  geom_hline(yintercept=0)
```



In our plot of residuals versus fitted values, we observe that the residuals are biased to fall below the line. This means there is not equal variance, which further suggests a lack of linearity between the two variables. These data suggest that the linear regression model does not accurately represent the relationship between our two variables, and we can say there is no linear relationship between the two. To assess the implications of these data for my hypothesis, I called the summary of the lmmodel statistics. From these we observe that my test statistic yields 1.525. We see that our p-value for this statistic is 0.127, which is not small enough to conclude that our result could not have occurred were there no linear relationship between UG and DEBT_MDN. We can also see that the standard error of our residuals is 3628, which is quite large and suggests low predictive value for our linear model. This means, in summary, our study does not provide sufficient evidence to reject the null hypothesis. It is likely that there is no linear relationship between the undergraduate population of a college and the median amount of debt of a student who attended it:

$$H_0 : \beta_1 = 0$$

Inferential Statistics for Small Schools

In order for our regression model to be valid for this data, we have to see that the residuals of the regression model are normally distributed, all trials are independent, there is a linear relationship between our variables, and the data has equal standard variance above and below the line at all points along the line. As I have suggested above, it seems that the residuals of our model are not normally distributed, there is not equal variance above and below our regression line, and there appears to be no linear relationship between our variables of interest. These facts in total suggest that our analysis is built on quite a shaky foundation, and probably is not useful to characterize our data. That being said, in order to flex my statistical analysis muscles I decided to construct an interval to see what my model would predict for the median debt amount of a school with 1500 undergraduate students, approximately how many students are at Whitman.

```r
newx <-data.frame(UG = 1500)

ci <-predict.lm(lmmodel, newx, interval = "confidence", level = 0.95)
ci
```

```
##        fit      lwr      upr
## 1 6840.147 6632.417 7047.878
```

From this prediction interval of the median debt amount of schools, we see that we are 95% confident that the actual median amount of debt of all students at a college in the United States of undergraduate population 1500 is between 6632.417 and 7047.878 USD. This would be my model's prediction for the median debt amount of students at Whitman.

Discussion of Error

Given that this dataset is quite large, containing 46 percent of all undergraduate students and 100% of all universities in the United States, the sample size of the data is sufficient to make the possibility of a Type II error quite small. With the large sample size in mind as well as the middling significance level of $\alpha = 0.05$, it is also unlikely that I am encountering a Type I error.

Conclusions:

We know that the conditions for the linear regression model were not met by our dataset. We also know that the dataset only included students who received federal financial aid, and well as only data on undergraduate population from the academic year 2000-2001. This means our data is quite weak and should not be taken as representative of the financial conditions of college graduates or undergraduate schools in the U.S. today. However, had we met all the necessary conditions, and had the data included all students from all years, our test statistic p-value would be interpreted as too low to suggest a correlation at all between undergraduate population size and median debt amount of a student from colleges in the U.S. This would have been a satisfying negative result to maintain the null hypothesis:

$$H_0 : \beta_1 = 0$$

Citations:

U.S. Dept. of Education (2015): National College Scorecard. U.S. Dept. of Education. https://collegescorecard.ed.gov

Rothwell, Jonathan (2015): Understanding the College Scorecard. Brookings Institution. https://www.brookings.edu/opinions/understanding-the-college-scorecard/